

ABSTRACT OF THE DISCLOSURE

Linguistic frequency data is encoded by identifying a plurality of sets of character strings in a source text, where each set comprises at least a first and a second character string. Frequency data is obtained for each set and stored at a

5 memory position in a first memory array that is assigned to each first character string. A pointer pointing to a position in the first memory array that has been assigned to the corresponding first character string of the respective set and which has stored the frequency data of the respective set, is stored in a second memory array for each set comprising each character string that is a second character string. The encoded data is accessed by identifying regions in the memory arrays that are each assigned a search string and a pointer pointing to a position in the first memory array.

10